16

Capturing and Training Motor Skills

Otniel Portillo-Rodriguez[1,2], Oscar O. Sandoval-Gonzalez[1],
Carlo Avizzano[1], Emanuele Ruffaldi[1] and Massimo Bergamasco[1]
[1]*Perceptual Robotics Laboratory, Scuola Superiore Sant'Anna, Pisa,*
[2]*Facultad de Ingeniería, Universidad Autonóma del Estado de México, Toluca,*
[1]*Italy*
[2]*México*

1. Introduction

Skill has many meanings, as there are many talents: its origin comes from the late Old English scele, meaning knowledge, and from Old Norse skil (discernment, knowledge), even if a general definition of skill can be given as "the learned ability to do a process well" (McCullough, 1999) or as the acquired ability to successfully perform a specific task.
Task is the elementary unit of goal directed behaviour (Gopher, 2004) and is also a fundamental concept -strictly connected to "skill"- in the study of human behaviour, so that psychology may be defined as the science of people performing tasks. Moreover skill is not associated only to knowledge, but also to technology, since technology is -literally in the Greek- the study of skill.
Skill-based behaviour represents sensory-motor performance during activities following a statement of an intention and taking place without conscious control as smooth, automated and highly integrated patterns of behaviour. As it is shown in Figure 1, a schematic representation of the cognitive-sensory-motor integration required by a skill performance, complex skills can involve both gesture and sensory-motor abilities, but also high level cognitive functions, such as procedural (e.g. how to do something) and decision and judgement (e.g. when to do what) abilities. In most skilled sensory-motor tasks, the body acts as a multivariable continuous control system synchronizing movements with the behavioural of the environment (Annelise Mark Pejtersen, 1997). This way of acting is also named also as, action-centred, enactive, reflection-in-action or simply know-how.
Skills differ from talent since talent seems native, and concepts come from schooling, while skill is learned by doing (McCullough, 1999). It is acquired by demonstration and sharpened by practice. Skill is moreover participatory, and this basis makes it durable: any teacher knows that active participation is the way to retainable knowledge.
The knowledge achieved by an artisan throughout his/her lifelong activity of work is a good example of a skill that is difficult to transfer to another person. At present the knowledge of a specific craftsmanship is lost when the skilled worker ends his/her working activity or when other physical impairments force him/her to give up. The above considerations are valid not only in the framework of craftsmanship but also for more general application domains, such as the industrial field, e.g. for maintenance of complex mechanical parts, surgery training and so on.

Source: Human-Robot Interaction, Book edited by: Daisuke Chugo,
ISBN 978-953-307-051-3, pp. 288, February 2010, INTECH, Croatia, downloaded from SCIYO.COM

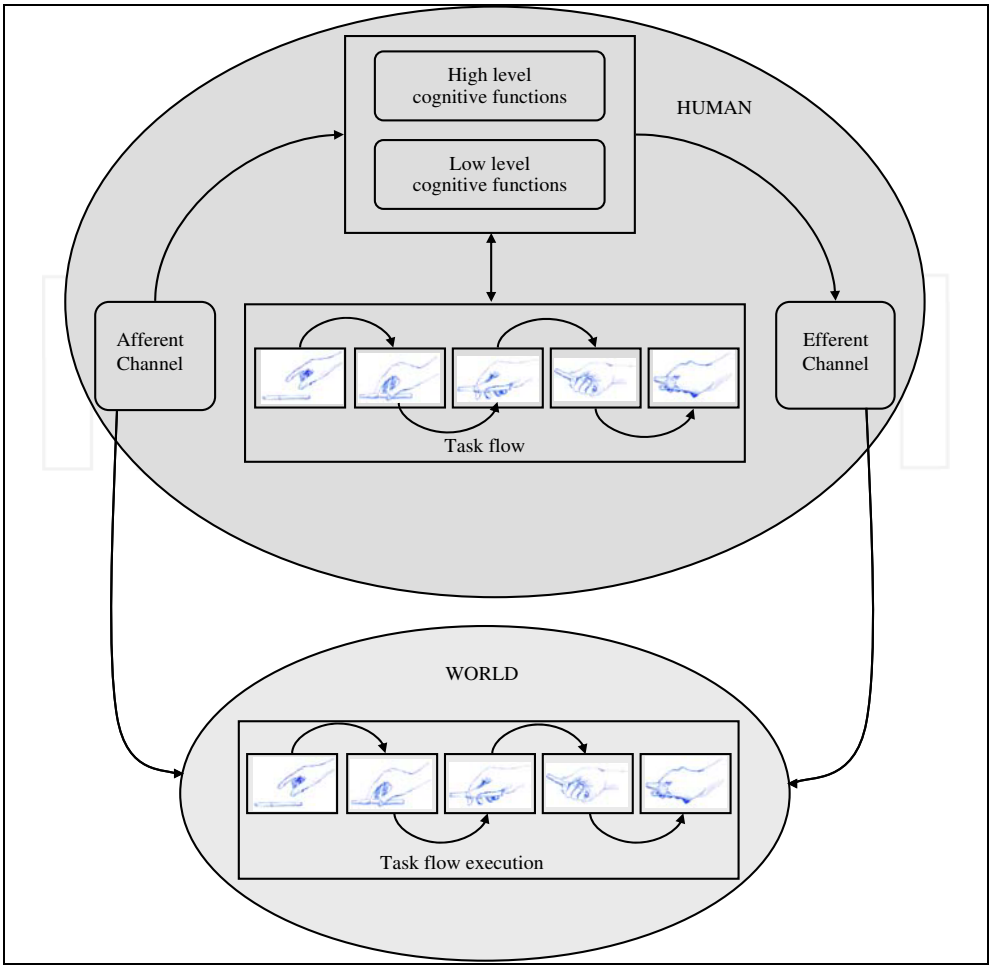

Fig. 1. A schematic representation of the cognitive-sensory-motor integration required by a skill performance

The research done stems out from the recognition that technology is a dominant ecology in our world and that nowadays a great deal of human behaviour is augmented by technology. Multimodal Human-Computer Interfaces aim at coordinating several intuitive input modalities (e.g. the user's speech and gestures) and several intuitive output modalities.

The existing level of technology in the HCI field is very high and mature, so that technological constrains can be removed from the design process to shift the focus on the real user's needs, as it is demonstrated by the fact that nowadays the user-centered design has became fundamental for devising successful everyday new products and interfaces. (Norman, 1986; Norman, 1988), fitting people and that really conforming their needs.

However, until now most interaction technologies have emphasized more input channel (afferent channel in Figure 1 The role of HCI in the performance of a skill), rather than output (efferent channel); foreground tasks rather than background contexts.

Advances in HCI technology allows now to have better gestures, more sensing combinations and improve 3D frameworks, and so it is possible now to put also more emphasis on the output channel, e.g. recent developments of haptic interfaces and tactile effector technologies. This is sufficient to bring in the actual context new and better instruments and interfaces for doing better what you can do, and to teach you how to do something well: so interfaces supporting and augmenting your skills. In fact user interfaces to advanced augmenting technologies are the successors to simpler interfaces that have existed between people and their artefacts for thousands of years (M. Chignell & Takeshit, 1999).

The objectives is to develop new HCI technologies and devise new usages of existing ones to support people during the execution of complex tasks, help them to do things well or better, and make them more skilful in the execution of activities, overall augmenting the capability of human action and performance.

We aimed to investigate the transfer of skills defined as the use of knowledge or skill acquired in one situation in the performance of a new, novel task, and its reproducibility by means of VEs and HCI technologies, using actual and new technology with a complete innovative approach, in order to develop and evaluating interfaces for doing better in the context of a specific task.

Figure 2 draws on the scheme of Figure 1, and shows the important role that new interfaces will play and their features. They should possess the following functionalities:

- Capability of interfacing with the world, in order to get a comprehension of the status of the world;
- Capability of getting input from the humans through his efferent channel, in a way not disturbing the human from the execution of the main task (transparency);
- Local intelligence, that is the capability of having an internal and efficient representation of the task flow, correlating the task flow with the status of the environment during the human-world interaction process, understanding and predicting the current human status and behaviour, formulating precise indications on next steps of the task flow or corrective actions to be implemented;
- Capability of sending both information and action consequences in output towards the human, through his/her afferent channel, in a way that is not disturbing the human from the execution of the main task.

We desire improving both input and output modalities of interfaces, and on the interplay between the two, with interfaces in the loop of decision and action (Flach, 1994) in strictly connection with human, as it is shown clearly in Figure 2. The interfaces will boost the capabilities of the afferent-efferent channel of humans, the exchange of information with the world, and the performance of undertaken actions, acting in synergy with the sensory-motor loop.

Interfaces will be technologically invisible at their best –not to decrease the human performance-, and capable of understanding the user intentions, current behaviour and purpose, contextualized in the task.

In this chapter a multimodal interface capable to understand and correct in real-time hand/arm movements through the integration of audio-visual-tactile technologies is presented. Two applications were developed for this interface. In the first one, the interface acts like a translator of the meaning of the Indian Dance movements, in the second one the interface acts like a virtual teacher that transfers the knowledge of five Tai-Chi movements

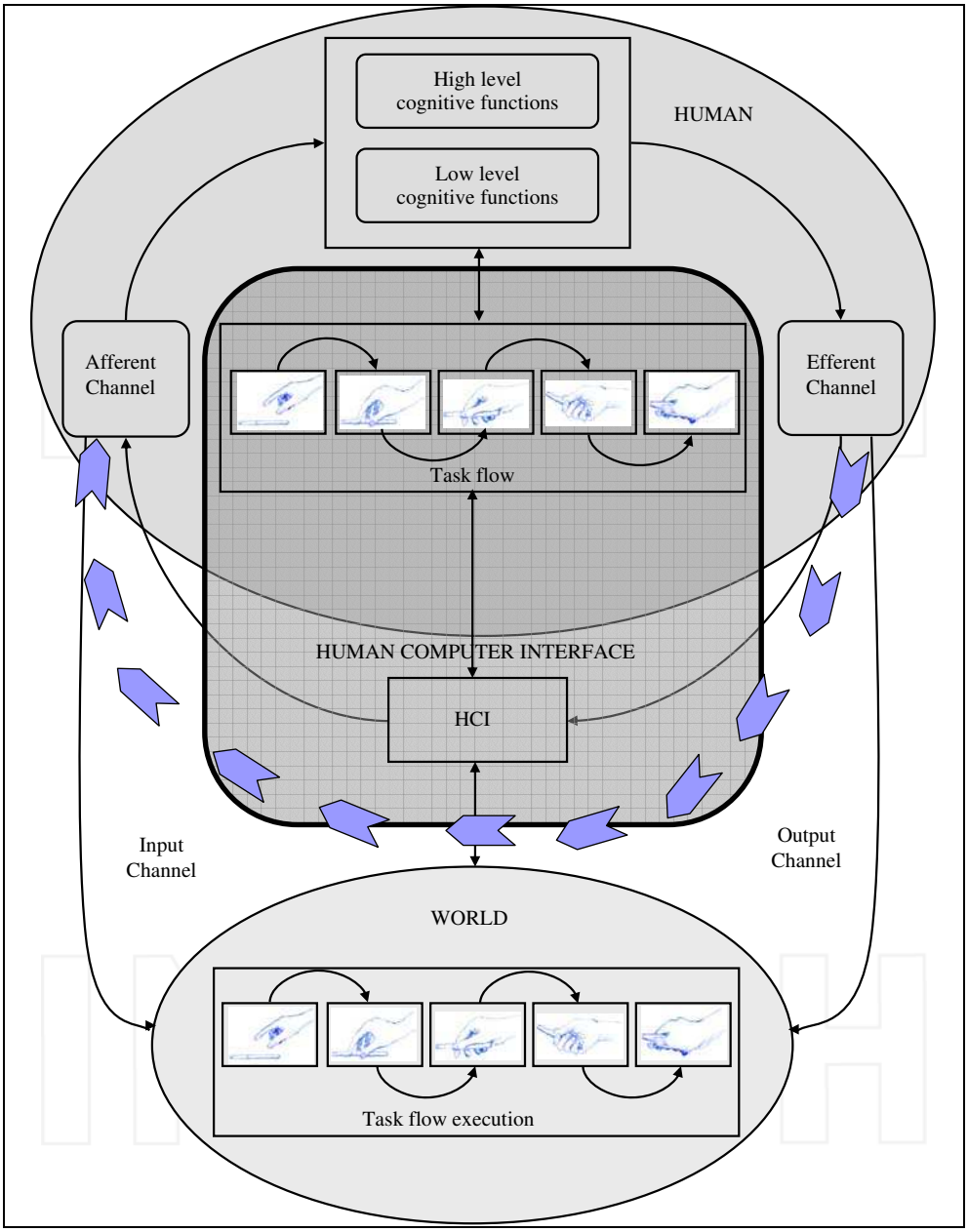

Fig. 2. The role of HCI in the performance of a skill

using feed-back stimuli to compensate the errors committed by a user during the performance of the gesture (Tai-Chi was chose due its movements must be performed precise and slow).

In both applications, a gesture recognition system is its fundamental component, it was developed using different techniques such as: k-means clustering, Probabilistic Neural Networks (PNN) and Finite State Machines (FSM). In order to obtain the errors and qualify the actual movements performed by the student respect to the movements performed by the master, a real-time descriptor of motion was developed. Also, the descriptor generate the appropriate audio-visual-tactile feedbacks stimuli to compensate the users' movements in real-time. The experiments of this multimodal platform have confirmed that the quality of the movements performed by the students is improved significantly.

2. Methodology to recognize 3D gestures using the state based approach

For human activity or recognition of dynamic gestures, most efforts have been concentrated on using state-space approaches (Bobick & Wilson, 1995) to understand the human motion sequences. Each posture state (static gesture) is defined as a state. These states are connected by certain probabilities. Any motion sequence as a composition of these static poses is considered a walking path going through various states. Cumulative probabilities are associated to each path, and the maximum value is selected as the criterion for classification of activities. Under such a scenario, duration of motion is no longer an issue because each state can repeatedly visit itself. However, approaches using these methods usually need intrinsic nonlinear models and do not have closed-form solutions. Nonlinear modeling also requires searching for a global optimum in the training process and a relative complex computing. Meanwhile, selecting the proper number of states and dimension of the feature vector to avoid "underfitting" or "overfitting" remains an issue.

State space models have been widely used to predict, estimate, and detect signals over a large variety of applications. One representative model is perhaps the HMM, which is a probabilistic technique for the study of discrete time series. HMMs have been very popular in speech recognition, but only recently they have been adopted for recognition of human motion sequences in computer vision (Yamato et al., 1992). HMMs are trained on data that are temporally aligned. Given a new gesture, HMM use dynamic programming to recognize the observation sequence (Bellman, 2003).

The advantage of a state approach is that it doesn't need a large set of data in order to train the model. Bobick (Bobick, 1997) proposed an approach that models a gesture as a sequence of states in a configuration space. The training gesture data is first manually segmented and temporally aligned. A prototype curve is used to represent the data, and is parameterized according to a manually chosen arc length. Each segment of the prototype is used to define a fuzzy state, representing transversal through that phase of the gesture. Recognition is done by using dynamic programming technique to compute the average combined membership for a gesture.

Learning and recognizing 3D gestures is difficult since the position of data sampled from the trajectory of any given gesture varies from instance to instance. There are many reasons for this, such as sampling frequency, tracking errors or noise, and, most notably, human variation in performing the gesture, both temporally and spatially. Many conventional gesture-modeling techniques require labor-intensive data segmentation and alignment work.

The attempt of our methodology is develop a useful technique to segment and align data automatically, without involving exhaustive manual labor, at the same time, the representation used by our methodology captures the variance of gestures in spatial-

temporal space, encapsulating only the key aspect of the gesture and discarding the intrinsic variability to each person's movements. Recognition and generalization is spanned from very small dataset, we have asked to the expert to reproduce just five examples of each gesture to be recognized.

As mentioned before, the principal problem to model a gesture using the state based approach is the characterization of the optimal number of states and the establishment of their boundaries. For each gesture, the training data is obtained concatening the data of its five demonstrations. To define the number of states and their coarse spatial parameters we have used dynamic k-means clustering on the training data of the gesture without temporal information (Jain et al., 1999). The temporal information from the segmented data is added to the states and finally the spatial information is updated. This produces the state sequence that represents the gesture. The analysis and recognition of this sequence is performed using a simple Finite State Machine (FSM), instead of use complex transitions conditions as in (Hong et al., 2000), the transitions depend only of the correct sequence of states for the gesture to be recognized and eventually of time restrictions i.e., minimum and maximum time permitted in a given state.

For each gesture to be recognized, one PNN is create to evaluate which is the nearest state (centroid in the configuration state) to the current input vector that represents the user's body position. The input layer has the same number of neurons as the feature vector (Section 3) and the second layer has the same quantity of hidden neurons as states have the gesture. The main idea is to use the states' centroids obtained from the dynamic k-means as weights in its correspondent hidden neuron, in a parallel way where all the hidden neurons computes the similarities of the current student position and its corresponding state. In our architecture, each class node is connected just to one hidden neuron and the number of states in which the gesture is described defines the quantity of class nodes. Finally, the last layer, a decision network computes the class (state) with the highest summed activation. A general diagram of this architecture is presented in the Figure 3.

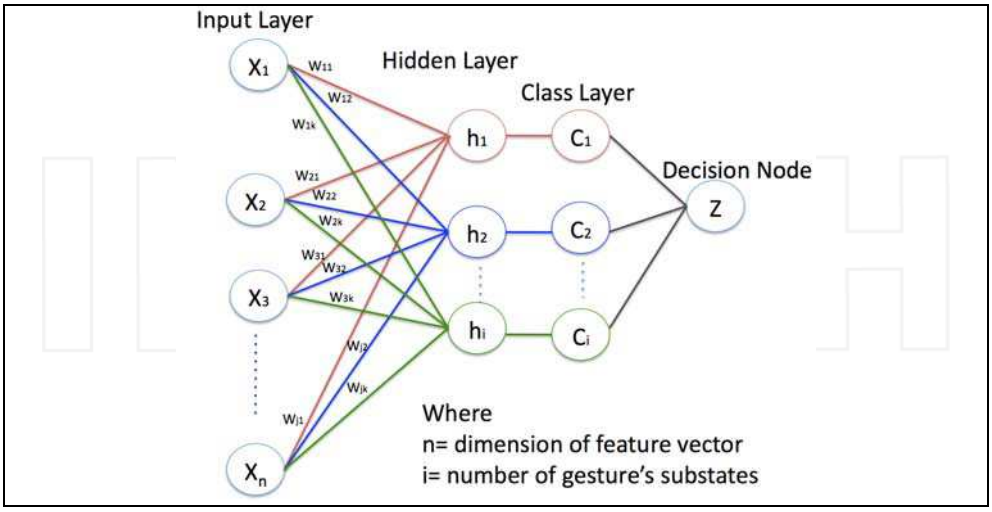

Fig. 3. PNN architecture used to estimate the most similar gesture's state from the current user's body position.

This approach allows real-time recognition while avoiding the classical disadvantages of this network: big computational resources (storage and time) during execution than many other models. An alternative for such computation is the use of RNN. In (Inamura et al., 2002) the authors tested the use of RNNs for motion recognition, according to their results, more than 500 nodes and more than 200,000 weight parameters between each node are needed in order to integrate the memorization process in the RNN. The RNN consist of motion elements neurons, symbol representation neurons and buffer neurons for treating time-series data. The required number of weights increases in proportion to the square of the number of all nodes. On the contrary in our methodology the number of parameters is proportional to the product of the number of hidden nodes (states in the gesture) and the dimension of the feature vector. To give a concrete example, a gesture typically has 10 states and the dimension of the feature vector is 13 (Section 3), resulting that with only 130 parameters a gesture is modeled, given as result a high information compression ratio.
The creation of the finite state machine is fast and simple; the transitions depend only of the correct sequence of states for the gesture to be recognized and eventually of time restrictions. If the FSM reaches its final state then the algorithm concludes that a gesture is recognized.

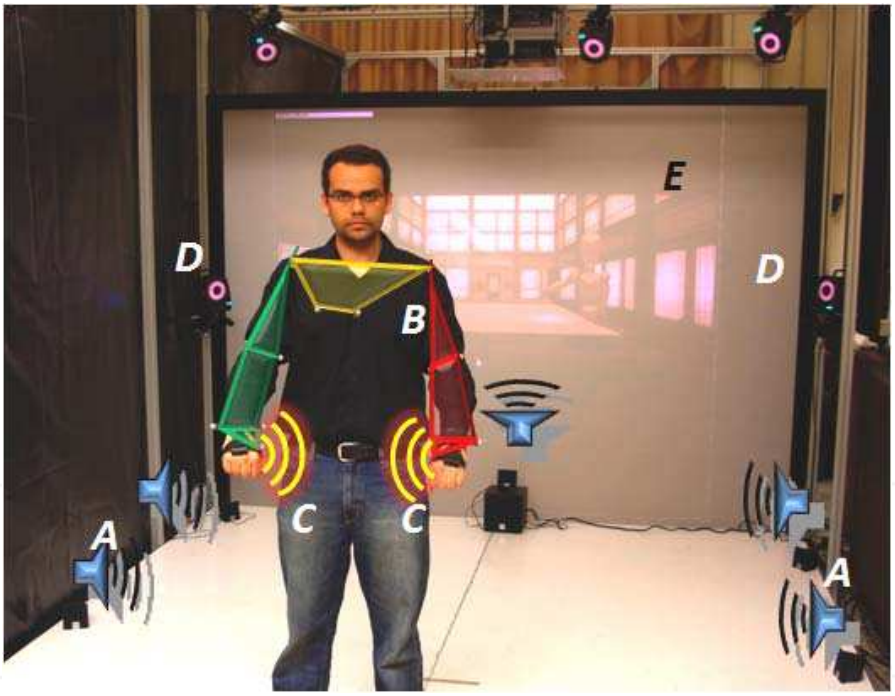

Fig. 4. Multimodal Platform set up, A) 3D sound, B) Kinematics Body C) Vibrotactile device (SHAKE) D) Vicon System E) Virtual Environment

3. Architecture of the multimodal interface

The hardware architecture of our Interface is composed of five components: VICON capture system (VCS), a host PC (with processor of 3 Ghz Intel Core Duo and 2 gigabytes of RAM

memory), one video projector, a pair of wireless vibrotactile stimulators and a 5.1 sound system. In the host PC, four applications run in parallel: Vicon Nexus (VICON, 2008), Matlab Simulink® and XVR (VRMedia, 2008). Depending of the application some of the components are not used.

The user has to perform a movement inside the capture system's workspace, depending of the application it could be an Indian Dance movement or a Tai Chi movement. Both applications where built using almost the same components, and uses the same methodology to create the recognizer blocks (pair of PNN-FSM) described in previous section. The Indian Dance application is an interactive demo in which five basic movements of the Indian dance can be recognized and their meaning can be represented using images and sound, its scope is more artistic than interactive, its development was done to test our methodology to recognize complex gestures.

In the Tai Chi application the objective was create a multimodal transfer skill system of Tai Chi movements, in this case, the gesture to be performed is known, the key idea is understand what far the current position of the user limbs is respect with the ideal position inside the reference movement. Once the distance (error) is calculated, it is possible modulate feedback stimuli (vision, audio and tactile) to indicate to the user the correct configuration of the upper limbs. It is important mention that the feature variables for gesture recognition and for gesture error are not the same.

It is possible observed from the Figures 8 and 11 that the acquisition data and recognition system (enclosed by a blue frame) in both applications are composed for the same blocks. For this reason in the next sections all the common elements in both applications there will be described, then each application it will be presented.

3. Modeling the upper limbs of the body

In order to track the 3D position of the markers using the VICON it is necessary create a kinematics model of the user's upper limbs. The model used is shown in Figure 5; it is composed of 13 markers united for hinge and balls joints. When the user is inside of the workspace of tracking system, it searches the correspondence between the model and the tracked markers, if a match exists; VICON sends the 3D position of all markers to Simulink via UDP.

The tracked positions by themselves are not useful in information for recognition due they are dependent of the position of the user in the capture system's workspace. We have chosen another representation of these elements that are invariant to the position and orientation of the user, allowing having enough degrees of freedom to model the movements of the user without over-fitting.

The 13 elements chosen are:

- Right & left elbows angles
- Right & left wrists pitch angles
- Right & left Shoulders angles
- The magnitude of the distance between palms
- Three spatial vector components from the back to the right palm
- Three spatial vector components from the back to the left palm

In order to recognized movements from different persons, it is necessary normalize the data to the "pattern user". The normalization relies in the fact that in the gestures to be

recognized, the ensemble of angles of the feature vector have approximately the same temporal behavior and their range of values are similar for different users independently of their body sizes due their arms can be seen as kinematics chains with the same joint variables with different lengths. The key idea is normalize just the components of the feature vector that involves distances. The normalize factor is obtained each time that a new user interact with the system measuring the length of his/her right arm.

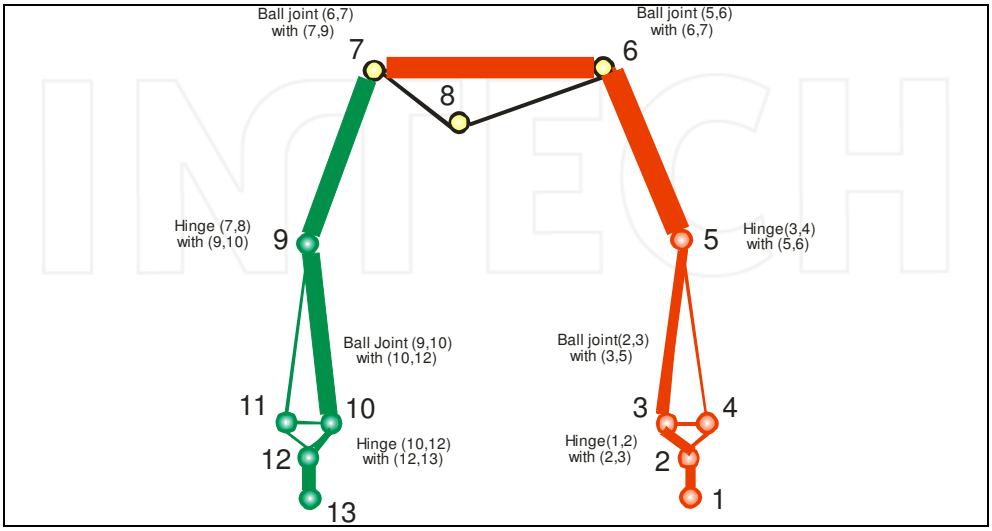

Fig. 5. Kinematics for the upper limbs based on the marker placement on the arms and hands

4. Cleaning and autocalibration algorithms

The motion of the user is tracked with the eight cameras and the electronics acquisition unit of the VICON system at 300 Hz. Sometimes, due to the markers obstructions in the human motion, the data information is lost. For this reason, the "cleaning algorithm" described in (Qian 2004), was implemented.

A calibration process was implemented in order to identify the actual position of the markers and adjust the kinematics model to the new values. Therefore, a fast (1ms) auto calibration process was designed in order to obtain the initial position of the markers of a person placed in a military position called "stand at attention". The algorithm checks the dimension of his/her arms and the position of the markers. The angles are computed and finally this information is compared with to the ideal values in order to compensate and normalize the whole system.

5. Capturing and recognizing Indian Dance movements in real time

In order to test our methodology to recognize gestures, we have implemented a system that recognizes seven basic movements (temporal gestures) of the Indian Dance. Each one has a meaning; thus, the scope of our system is to discover if the user/dancer has performed a valid known movement in order to translate their meaning in an artistic representation using graphics and sounds.

The gestures to be recognized are: earth, fish, fire, sky, king, river and female. In Figure 6, different phases of each gesture are presented. The gestures were chosen from the Indian Dance and their spatial-temporal complexity was useful to test our gesture recognition methodology.

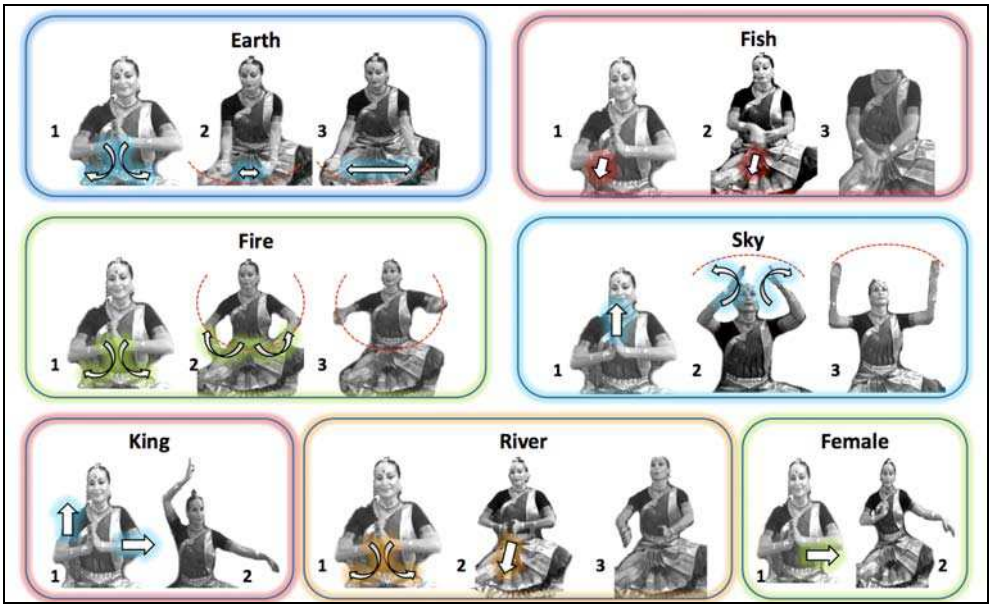

Fig. 6. The seven Indian Dance movements that the system recognizes

The interaction with the user is simple and easy to learn (Figure 7). At the beginning the user remains in front of the principal screen in a motionless state. Once the system has sensed the user's inactivity, it sends a message indicating that it is ready to capture the movement of the user. If the system recognizes the movement, it renders a sound and image corresponding to its meaning. After that, the system is again ready to capture a new movement.

It is possible observe how the cameras track the user movements using the reflective markers that are mounted on the suit dressed by the user. The user's avatar and the image that represents the meaning (a King) of the recognized gesture are displayed in the screen.

The operation of the recognition system is as follow (it is useful see the Figure 8 to check the flow of the information): The eight cameras and their electronics acquisition that compose the VCS acquire the 3D positions of reflective markers attached to a suit that the user dresses on its upper limbs and send the positions to Simulink through UDP protocol.

In Matlab Simulink's Real-Time Windows Target, we have developed a real-time recognition system with a sample rate of 50 Hz. Its operation at each frame rate it's as follows: the current 3D position of the markers is read from Nexus, then the data are filtered to avoid false inputs, then, the data are send simultaneously to XVR (to render a virtual avatar) and other block that converts the 3D positions of the 13 markers (vector of 39 elements) in a normalized feature vector (values from 0 to 1). The elements of feature vector were described in the section 3.

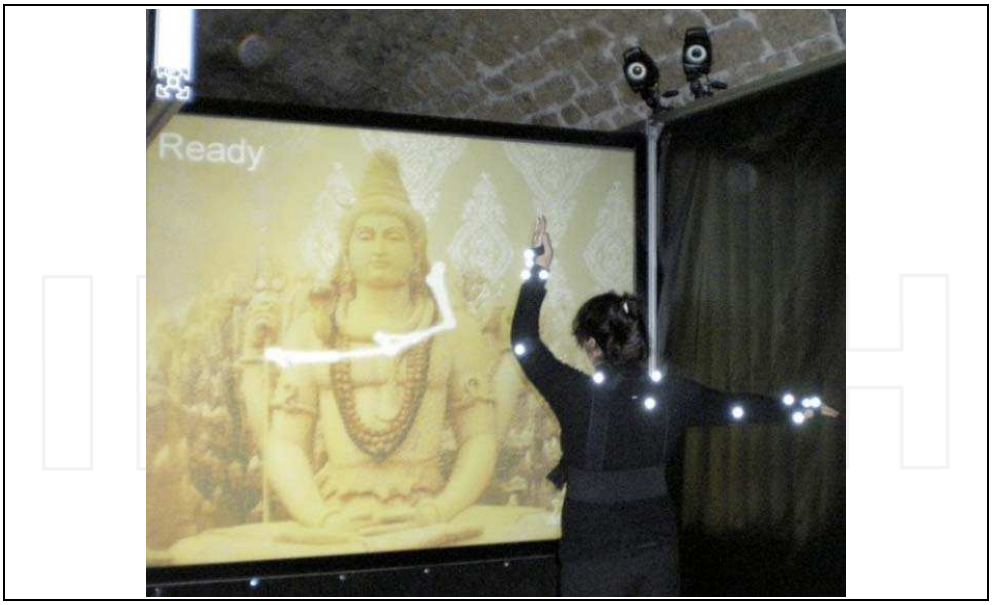

Fig. 7. The recognition system in action.

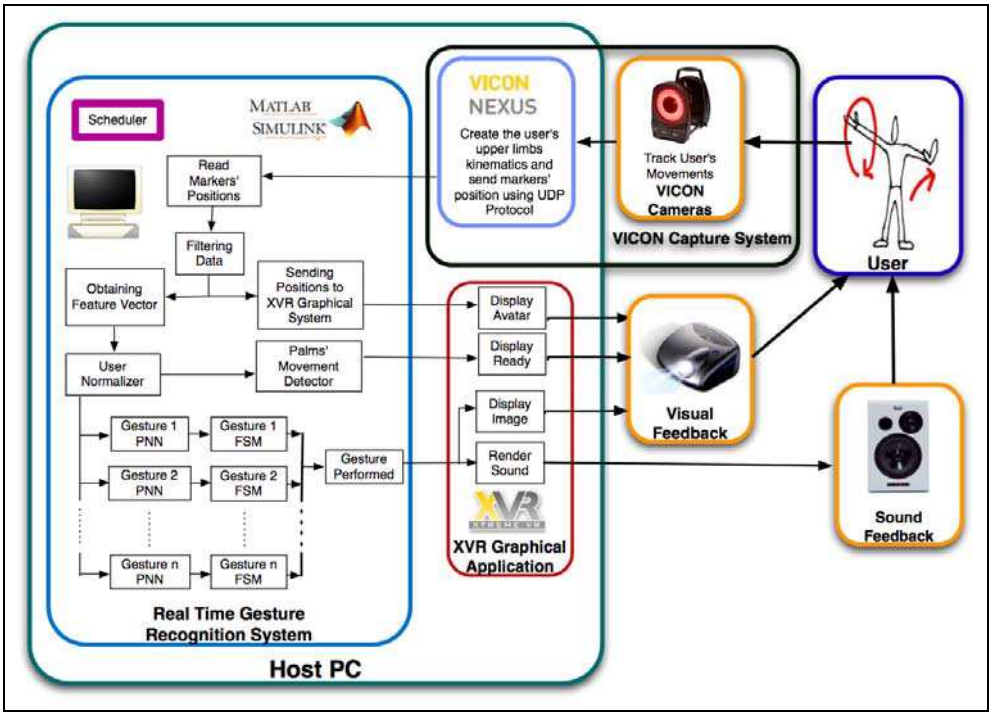

Fig. 8. Architecture of the Indian Dance Recognition System

Then, the normalized feature vector is sent simultaneously to a group of PNN–FSM couples that work in parallel. Each couple is used to recognize one gesture; the PNN is used to determinate which is the nearest state respect to the current body position for one particular gesture. Recognition is performed using a FSM, where, its state transitions depend only of the output of the its PNN. If the FSM arrives to its final state the gesture is recognized.
The movements of both palms are analyzed in order to determine if the user is or is not in motion, this information is useful in order to interact correctly with the user. All the FSM are initialized when no motion of the palms is detected.
XVR, a virtual environment development platform is used to display visual and sound information to the user. An avatar shows the users' movements, a silhouette is used to render the sensation of performing movements along the time. The graphical application also indicates when is ready to recognized a new movement. If the recognition system has detected a valid gesture, the meaning of the gesture recognized is displayed to the user using an appropriate image and sound (Figure 9).

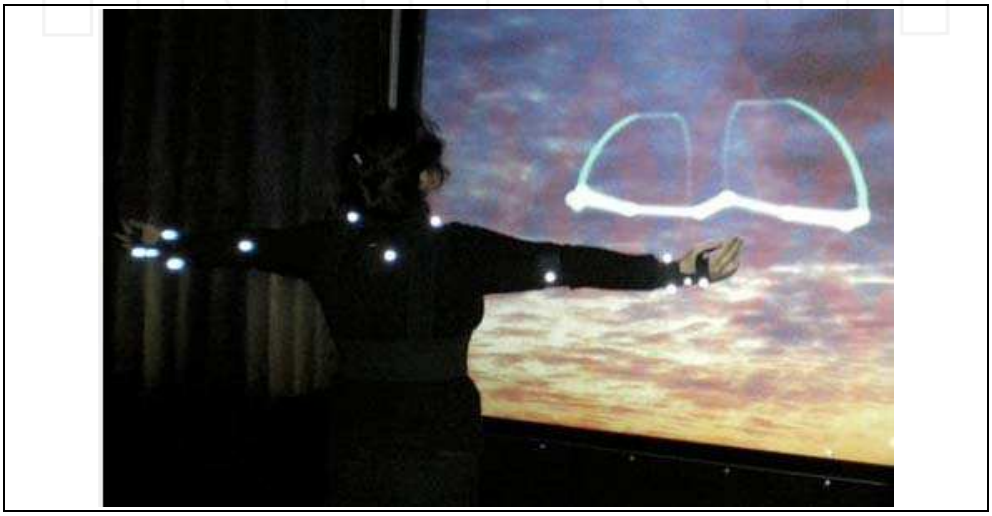

Fig. 9 .The gesture that represents "Sky" has been recognized. The user can see the avatar, hear the sound of a storm and see the picture of the firmament.

6. Development of a real-time gesture recognition, evaluation and feed-forward correction of a multimodal Tai-Chi platform

The learning process is one of the most important qualities of the human being. This quality gives us the capacity to memorize different kind of information and behaviors that help us to analyze and survive in our environment. Approaches to model learning have interested researches since long time, resulting in such a way in a considerable number of underlying representative theories.
One possible classification of learning distinguishes two major areas: Non-associative learning like habituation and sensitization, and the associative learning like the operant conditioning (reinforcement, punish and extinction), classical conditioning (Pavlov Experiment), the observational learning or imitation (based on the repetition of a observed

process) (Byrne & Russon, 1998), play (the perfect way where a human being can practice and improve different situations and actions in a secure environment) (Spitzer, 1988), and the multimodal learning (dual coding theory) (Viatt & Kuhn, 1997).
Undoubtedly, the imitation process has demonstrated a natural instinct action for the acquisition of knowledge that follows the learning process mentioned before. One example of multimodal interfaces using learning by imitation in Tai-chi has been applied by the Carnegie Mellon University in a Tai-Chi trainer platform (Tan, 2003), demonstrating how through the use of technology and imitation the learning process is accelerated.
The human being has a natural parallel multimodal communication and interaction perceived by our senses like vision, hearing, touch, smell and taste. For this reason, the concept of Human-Machine Interaction HMI is important because the capabilities of the human users can be extended and the process of learning through the integration of different senses is accelerated (Cole & Mariani, 1995; Sharma et al., 1998; Akay et al., 1998)). Normally, any system that pretends to have a normal interaction must be as natural as possible (Hauptmann & McAvinney, 1993). However, one of the biggest problems in the HMI is to reach the transparency during the Human-Machine technology integration.
In such a way, the multimodal interface should present information that answers to the "why, when and how" expectations of the user. For natural reasons exists a remarkable preference for the human to interact multimodally rather than unimodally. This preference is acquired depending of the degree of flexibility, expressiveness and control that the user feels when these multimodal platforms are performed (Oviatt, 1993). Normally, like in real life, a user can obtain diverse information observing the environment. Therefore, the Virtual Reality environment (VR) concept should be applied in order to carry out a good Human-Machine Interaction. Moreover, the motor learning skills of a person is improved when diverse visual feedback information and correction is applied (Bizzi et al., 1996).
For instance the tactile sensation, produced on the skin, is sensitive to many qualities of touch. (Lieberman & Breazeal, 2007) carried out, for first time, an experiment in real time with a vibrotactile feedback to compensate the movements and accelerate the human motion learning. The results demonstrate how the tactile feedback induces a very significant change in the performance of the user. In the same line of research Boolmfield performed a Virtual Training via Vibrotactile Arrays (Boolmfield & Badler, 2008).
Another important perception variable is the sound because this variable can extend the human perception in Virtual Environments. The modification of parameters like shape, tone and volume in the sound perceived by the human ear (Hollander & Furness, 1994), is a good approach in the generation of the description and feedback information in the human motion.
Although a great grade of transparency and perception capabilities are transmitted in a multimodal platform, the intelligence of the system is, unquestionably, one of the key parts in the Human-Machine interaction and the transfer of a skill. Because of the integration, recognition and classification in real-time of diverse technologies are not easy tasks, a robust gesture recognition system is necessary in order to obtain a system capable to understand and classify what a user is doing and pretending to do.

6.1 Tai-Chi system implementation

This application teaches to novel students, five basic Tai Chi movements. Tai-Chi movements were chosen because they have to be performed slowly with high precision.

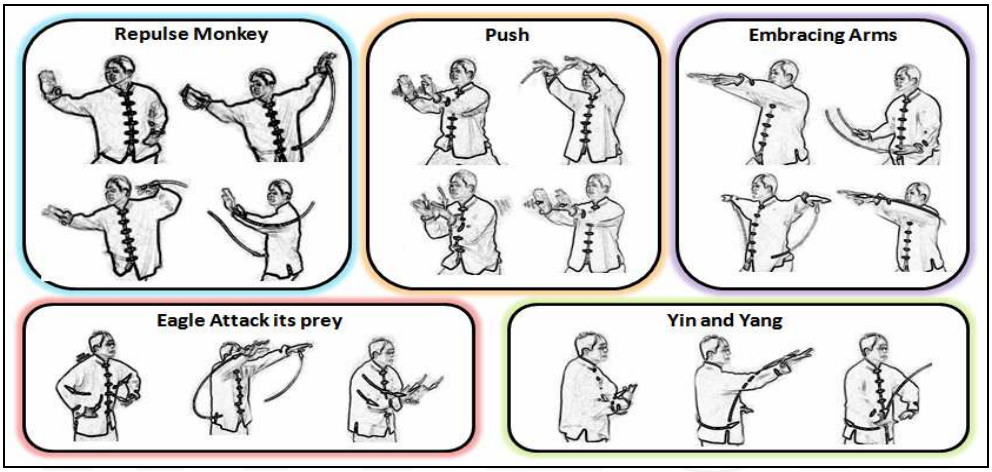

Fig. 10. The 5 Tai-Chi Movements

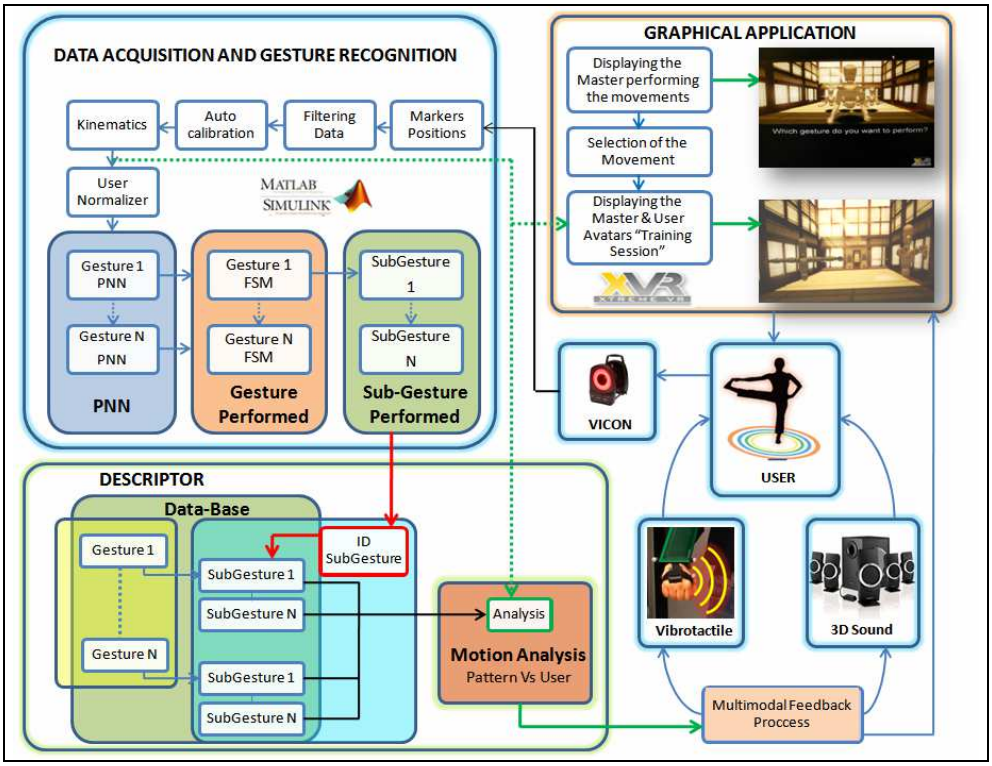

Fig. 11. Architecture of the Multimodal Tai Chi Platform System

Each movement is identified and analyzed in real time by the gesture recognition system. The gestures performed by the users are subdivided in n-spatial states (time-independent)

and evaluated step-by-step in real time by the descriptor system. Finally, the descriptor executes audio-visual-tactile feedback stimuli in order to try to correct the user's movements. The general architecture of the multimodal platform is shown in Figure 11.
The data acquisition and gesture recognition blocks of this application practically are identical to the Indian Dance Application. The PNNs and FSMs used to recognize the gestures were changed to recognize the new movements. A new virtual environment was developed and in this application a vibrotactile feedback in addition to the visual and auditive was employed.

6.1.1 Real-time descriptor process

The comparison and qualification in real-time of the movements performed by the user is computed by the descriptor system. In other words, the descriptor analyzes the differences between the movements executed by the expert and the movement executed by the student, obtaining the error values and generating the feedback stimuli to correct the movement of the user.
The Real-Time Descriptor is formed by a database where for each gesture n instances of 26 variables (where n it is number of states of the gesture) with the following information are saved:

- Right & left elbows angles
- Right & left pitch and yaw wrists angles
- Right & left pitch, yaw and roll shoulders angles
- The magnitude of the distance between palms
- The magnitude of the distance between elbows
- Three spatial vector components from the back to the right palm
- Three spatial vector components from the back to the left elbow
- Three spatial vector components from the back to the right elbow

The descriptor's database is created through an offline process as follow:

1. For each gesture it is necessary capture from a pattern movement performed by the expert samples of the 26 mentioned variables
2. Each sample must be classified in its corresponding state normalizing the 13 variables described in the section 3 and feeding them to their corresponding PNN. Once that all the samples were classified, the result it is a sequence of estates to which each sample belongs
3. The sequence's indexes of the n-1 transitions between states plus the last index that conform the sequence are detected
4. With the n indexes founded, their corresponding data in the original samples are extracted to conform the data used to describe the gesture
5. Change of gesture and repeat the steps 1-5 until finish all the gestures
6. Create the descriptor database of all gestures

When the application is running, each state or subgesture is recognized in real time by the gesture recognition system during the performance of the movement. Using the classic feedback control loop during the experiments was observed that the user feels a delay in the corrections. For that reason, a feed-forward strategy was selected to compensate this perception. In this methodology when a user arrives at one state of the gesture, the descriptor system carries out an interpolation process to compare the actual values with respect to the values in the descriptor for the following state, creating a feed-forward loop which estimates in advance the next correction values of the movement. The error is computed by:

$$\theta_{error} = [P_{n+1} - U_n] * F_n \quad (1)$$

Where θ_{error} is the difference between the pattern and the user, P is the pattern value obtained from the descriptor, U is the user value, F_{u} is the normalize factor and n is the actual state.

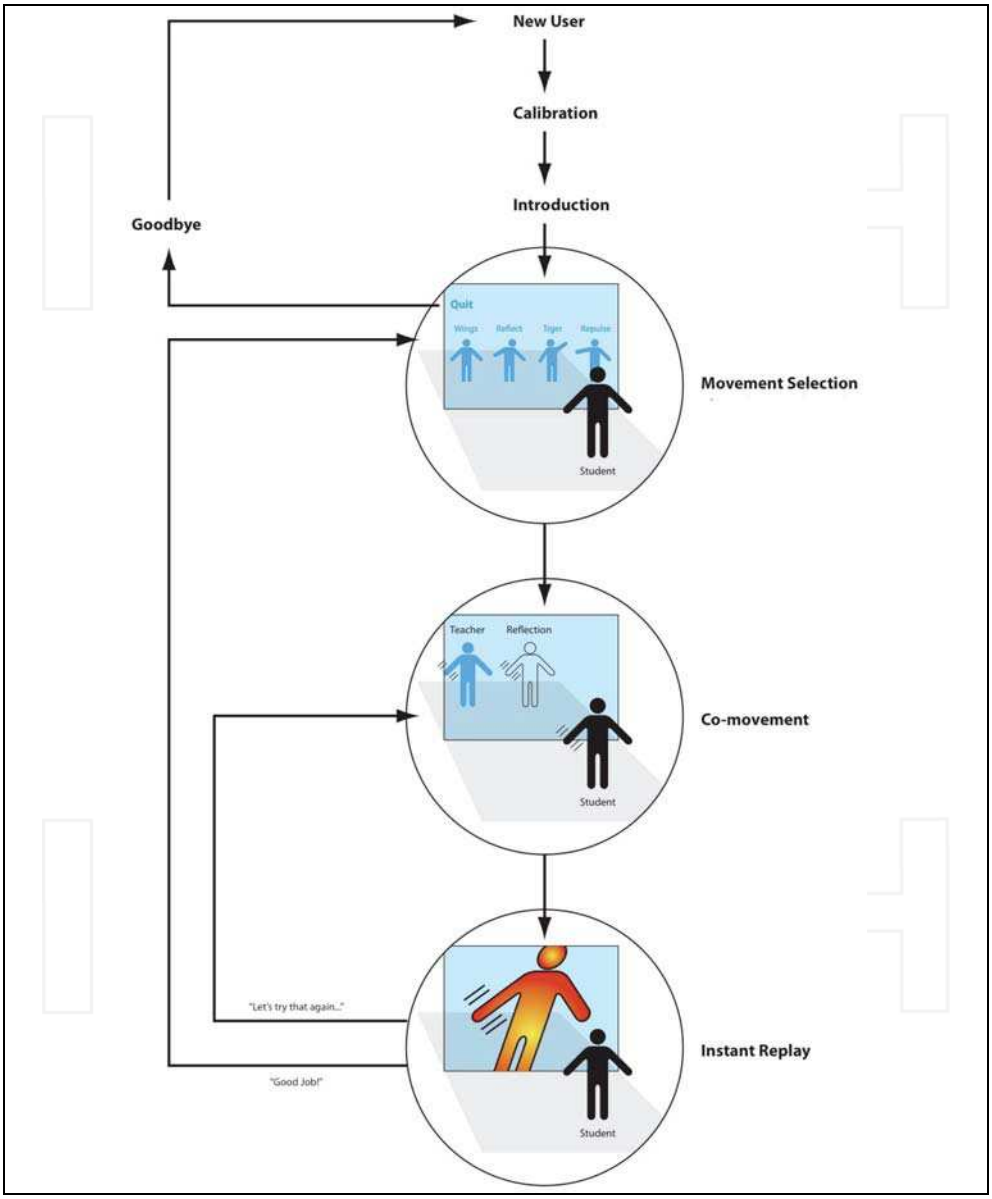

Fig. 12. Storyboard for the Multimodal Interface for Tai Chi Training

6.1.2 Virtual reality platform

The virtual environment platform provides the visual information to the user was programmed in XVR. There are 3 different sequences involved in this scenery. The first one is the initial screen that shows 5 avatars executing different Tai Chi movements. When a user tries to imitate one movement, the system recognizes the movement through the gesture recognition algorithm and passes the control to the second stage called "training session". In this part, the system visualizes 2 avatars, one represents the master and the other one is the user. Because learning strategy is based on the imitation process, the master performs the movement one step forward to the user. The teacher avatar remains in the state n+1 until the user has reached or performed the actual state n.

With this strategy the master gives the future movement to the user and the user tries to reach him. Moreover, the graphics displays a virtual energy line between the hands of the user. The intensity of this line is changing proportionally depending on the error produced by the distance between the hands of the student. When a certain number of repetitions have been performed, the system finishes the training stage and displays a replay session that shows all the movements performed by the student and the statistical information of the movement's performance. Figure 12 shows the storyboard for the interaction with the user and Figure 13 (A)(B) shows the virtual Tai-Chi environment.

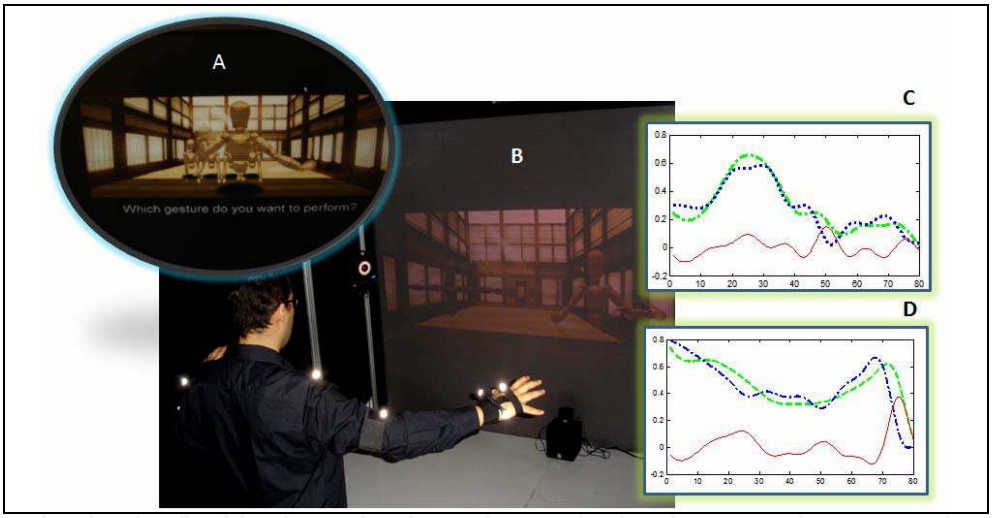

Fig. 13. VR environment, A) Initial Screen, 5 avatars performing Tai-Chi movements, B) Training session, two avatars, one is the master and second is the user. C) Distance of the Hands, D) Right Hand Position.

6.1.3 Vibrotactile feedback system

The SHAKE device was used to obtain wireless feedback vibrotactile stimulation. This device contains a small motor that produces vibrations at different frequencies. In this process, the descriptor obtains the information of the distance between the hands, after this, the data is compared with the pattern and finally sends a proportional value of the error. The SHAKE varies proportionally the intensity of the vibration according to error value

produced by the descriptor (1 Hz – 500 Hz). This constraint feedback is easy to understand for the users when the arms have reached a bad position and need to be corrected. Figure 13 (C) shows the ideal distance between the hands (green), the distance between the hands performed by the user (blue) and the feed-back correction (red).

6.1.4 Audio feedback system

The position of the arms in the X-Y plane is analyzed by the descriptor and the difference in position between the pattern and the actual movement in each state of the movement is computed. A commercial Creative SBS 5.1 audio system was used to render the sound through 5 speakers (2 Left, 2 Right, 1 Frontal) and 1 Subwoofer. In this platform was selected a background soft-repetitive sound with a certain level of volume. The sound strategy performs two major actions (volume and pitch) when the position of the hands exceeds the position of the pattern in one or both axes. The first one increases, proportionally to the error, the volume of the speakers in the corresponding axis-side (Left-Center-Right) where is found the deviation and decreases the volume proportionally in the rest of the speakers. The second strategy varies proportionally the pitch of the sound (100-10KHz) in the corresponding axis-side where was found the deviation. Finally, the user through the pitch and the volume can obtain information which indicates where is located the error and its intensity in the space.

7. Experimental results

The experiments were performed capturing the movements of 5 Tai-Chi gestures (Figure 10) from 5 different subjects. The tests were dived in 5 sections where the users performed 10 repetitions of the each one of the 5 movements performed. In the first section was avoid the use of technology and the users performs the movement in a traditional way, only observing a video of a professor performing one simple tai-chi movement. The total average error TAVG is calculated in the following way:

$$TAVG = \frac{1}{Ns}\sum_{s=0}^{Ns}\frac{1}{n}\sum_{i=0}^{n}(\theta_{Teacher} - \theta_{Student}) \quad (2)$$

Where N_s is the total number of subjects, n is the total number of states in the gesture and θ is the error between the teacher movement and the student.

Figure 14 (A) shows the ideal movements (Master Movements) of the gesture number 1 and (B) represents the TAVG of the gesture 1 executed by the 5 subjects without feedback. The TAVG value the 5 subjects without feedback was around 34.79% respect to the ideal movement. In the second stage of the experiments, the Virtual Reality Environment was activated. The TAVG value for the average of the 5 subjects in the visual feedback system presented in Figure 14(C) was around 25.31%. In the third section the Visual-Tactile system was activated and the TAVG value was around 15.49% respect to the ideal gesture. In the next stage of the experiments, the visual- 3D audio system was performed and the TAVG value for the 5 subjects in the audio-visual feedback system was around 18.42% respect to the ideal gesture. The final stage consists in the integration of the audio, vibrotactile and visual systems. The total mean error value for the average of the 5 subjects in the audio-visual-tactile feedback system was around 13.89% respect to the ideal gesture. Figure 14 (D) shows the results using the whole integration of the technologies.

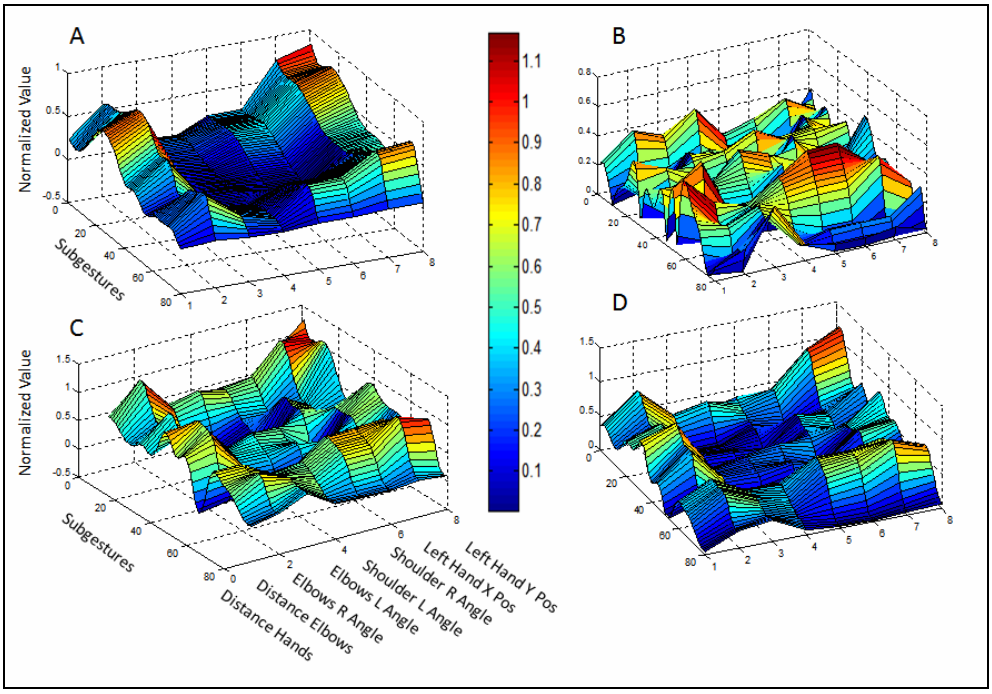

Fig. 14. Variables of Gesture 1, A) Pattern Movement, B) Movement without feedback, C) Movement with Visual feedback and D) Signals with Audio-Visual-Tactile feedback.

Figure 15 presents an interesting graph where the results of the four experiments are indicated. In one hand, as it was expected, the visual feedback presented the major error. In the other hand the integration of audio-visual-vibrotactile feedback has produced a significant reduction of the error of the users. The results of the experiments show that although the process of learning by imitation is really important, there is a remarkable improvement when the users perform the movements using the combination of diverse multimodal feedbacks systems.

8. Conclusion

We have built an intelligent multimodal interface to capture, understand and correct in real time a complex hand/arm gestures performed inside its workspace. The interface is formed by a commercial vision tracking system, a commercial PC and feedback devices: 3D sound system, a cave like VE and a pair of wireless vibrotactile devices. The interface can capture the upper part limbs kinematics of the user independently of the user's size and high. The interface recognizes complex gestures due a novel recognition methodology based on several machine-learning techniques such as: dynamic k-means, probabilistic neural networks and finite state machines. This methodology is the main contribution of this research Human Hand Computer Interaction research area, its working principle is simple: a gesture is split in several states (a state is an ensemble of variables that define an static position or configuration), the key is obtain the optimal number of states that define

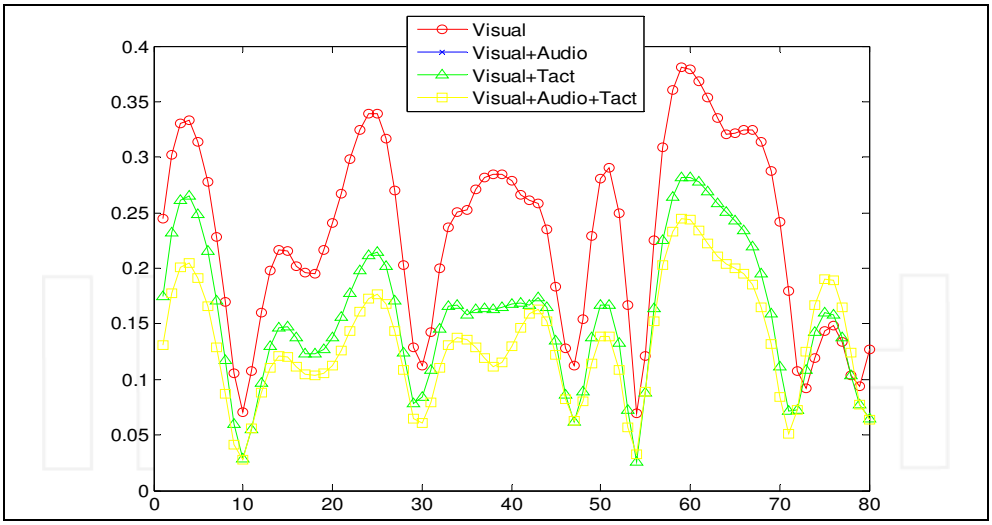

Fig. 15. Average Errors

correctly a gesture and develop an algorithm that recognize which is the most similar state to the current position of the user limbs; then the gesture recognition is simple due that just it is necessary check the sequence of states that the user generated with his/her movement, if the sequence is correct and arrives to the gesture's last state without error, the gesture is recognized.

The methodology proposed showed the effectiveness of dynamic k-means to obtain the optimal number and spatial position of each state. To calculate the boundaries of each state instead to use complex sequential algorithms such as Hidden Markov Models or recurrent neural networks, we have employed Probabilistic Neural Networks. For each gesture a PNN was created using as a hidden neurons the states founded by the dynamic k-means algorithm, this way a gesture can be modeled with few parameters enabling compress the information used to describe the gesture.

Furthermore the PNN is used not only to model the gesture but also to recognize it, avoiding use two algorithms. For example when a recognizer is developed with HMM its necessary at least executed two algorithms, the first one defines the parameters of the HMM given a dataset of sequences using the Baum-Welch algorithm and then, online the forwad-backward algorithm computes the probability of a particular output sequence and the probabilities of the hidden state values given that output sequence. This approach it is neither intuitive nor easy to implement when the sequence of data is multidimensional, to solve this problem, researchers that desire recognize complex gestures use dimension reductions algorithms (such as principal components analysis, independent components analysis or linear discriminant analysis) or transform the time dependent information to its frequential representation destroying their natural representation (positions, angles, distances, etc). Our methodology shown its effectiveness to recognize complex gesture using PNN with a feature vector of 16 dimensions without reduce its dimensionality.

The comparison and qualification in real-time of the movements performed by the user is computed by the descriptor system. In other words, the descriptor analyzes the differences

between the movements executed by the expert and the movements executed by the student, obtaining the error values and generating the feedback stimuli to correct the movements of the student. The descriptor can analyze step by step the movement of the user and creates a comparison between the movements by the master and user. This descriptor can compute the comparison up to 26 variables (angles, positions, distances, etc). For the Tai-Chi skill transfer system, only four variables were used which represents X-Y deviation of each hand with respect to the center of the body, these variables were used to generate spatial sound, vibrotactile and visual feedback. The study shown that with the use of this interface, the Tai Chi students improve to its capability to imitate their movements.

A lot of work must be done, first is still not clear the contribution of each feedback stimuli to correct the movements, seems that the visual stimuli (Master avatar) dominate to the auditive and vibrotactile feedbacks. A separate studies in which auditive and vibrotactile feeback will be the only stimulus must be done in order to understand their contributions to create the multimodal feeback. For the auditive study, a 3D spatial sound system must be developed putting emphasis in the Z position. For the vibrotactile study, a upper limbs suit with tactors distributed along the arm/hands must be developed, the position of the tactors must be studied through a psychophysical tests.

Once the multimodal platform has demonstrated the feasibility to perform the experiments related to the transfer of a skill in real-time, the next step will be focused in the implementation of a skill methodology which consists, in a brief description, into acquire the data from different experts, analyze their styles and the descriptions of the most relevant data performed in the movement and, through this information, select a certain lessons and exercises which can help the user to improve his/her movements. Finally it will be monitored these strategies in order to measure the progress of the user and evaluate the training. These information and strategies will help us to understand in detail the final effects and repercussions that produce each multimodal variable in the process of learning.

9. References

Akay, M., Marsic, I., & Medl, A. (1998). A System for Medical Consultation and Education Using Multimodal Human/Machine Communication. *IEEE Transactions on information technology in Biomedicine , 2*.

Annelise Mark Pejtersen, J. R. (1997). Ecological Information Systems and Support of Learning: Coupling Work Domain Information to user Characteristics. *Handbook of Human-Computer Interaction.* North/Holland.

Bellman, R. (2003). *Dynamic Programming*. Princeton University Press.

Bizzi, E., Mussa-Ivaldi, F., & Shadmehr, R. (1996). *Patent n. 5,554,033*. United States of America.

Bloomfield, A., & Badler, N. (2008). Virtual Training via vibrotactile arrays. *Teleoperator and Virtual Environments , 17*.

Bobick, A. F., & Wilson, A. D. (1995). A state-based technique for the summarization and recognition of gesture. *5th International Conference on Computer Vision*, (p. 382-388).

Bobick, A., & Davis, J. (1996). Real-Time Recognition of Activity Using Temporal Templates. *Proc. Int'l Conf. Automatic Face and Gesture Recognition.* Killington, Vt.

Byrne, R., & Russon, A. (1998). Learning by imitation: a Hierarchical Approach. *Behavioral and Brain Sciences , 21*, 667-721.

Cole, E., & Mariani, J. (1996). Multimodality. Survey of the State of the Art of Human Language Technolgy.

Flach, J. M. (1994). Beyond the servomechanism: Implications of closed-loop, adaptive couplings for modeling humna-machine systems. *Symposium on Human Interaction with Complex Systems.* North Carolina A&T State University.

Gopher, D. (2004). Control processes in the formation of task units. *28th International Congress of Psychology.* Beijing, China.

Hauptmann, A., & McAvinney, P. (1993). Gesture with Speech for Graphics Manipulation. *Man-Machines Studies , 38*.

Hollander, A., & Furness, T. A. (1994). Perception of Virtual Auditory Shapes. *Proceedings of the International Conference on Auditory Displays.*

Hong, P., Turk, M., & Huang, T. S. (2000). Gesture Modeling and Recognitition Using Finite State Machines. *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition.*

Inamura, T., Nakamura, Y., & Shimozaki, M. (2002). Associative Computational Model of Mirror Neurons that connects mising link betwwen behavior and symbols. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. .* Lausanne, Switzerland.

Jain, A., Murty, M. N., & Flynn, P. (1999). Data Clustering: A Review. *ACM Computing Surveys , 31* (3).

Lieberman, J., & Breazeal, C. (2007). Development of a wearable Vibrotactile FeedBack Suit for Accelerated Human Motor Learning. *IEEE International Conference on Robotics and Automation.*

McCullough, M. (1999). *Abstracting Craft.* MIT Press.

Norman, D. (1986). *User-centered systems design.* Hillsdale.

Norman, D. (1988). *The design of everyday things.* New York: Basic Books.

M. Chignell, P., & Takeshit, H. (1999). Human-Computer Interaction: The psychology of augmented human behavior. In P. Hancock (A cura di), *Human performance and Ergonomics.* Academic Press.

Oviatt, S. (1993). User Centered Modeling and Evaluation of Multimodal Interfaces. *Proceedings of the IEEE , 91*.

Qian, G. (2004). A gesture-Driven Multimodal Interactive Dance System. *IEEE International Conference on Multimedia and Expo.*

Sharma, R., Huang, T., & Pavlovic, V. (1998). A Multimodal Framework for Interacting With Virtual Environments. In C. Ntuen, & E. Park (A cura di), *Human Interaction With Complex Systems* (p. 53-71). Kluwer Academic Publishers.

Spitzer, M. (1998). *The mind within the net: models of learning, thinking and acting.* The MIT press.

Tan Chau, P. (2003). Training for physical Tasks in Virtual Environments: Tai Chi. *Proceedings of the IEEE Virtual Reality.*

Viatt, S., & Kuhn, K. (1997). Integration and synchronization of input modes during multimodal human-computer interaction. *Proc. Conf. Human Factors in Computing Systems CHI.*

VICON. (2008). Seen at December 29, 2008 from http://www.vicon.com

VRMedia. (2008). Seen at December 29, 2008 from EXtremeVR: virtual reality on the web: http://www.vrmedia.com

Yamato, J., Ohya, J., & Ishii, K. (1992). Recognizing human action in time-sequential images using Hidden Markov Model. *IEEE Conference CPVPR*, (p. 379-385). Champaign, IL.

Human-Robot Interaction

Edited by Daisuke Chugo

ISBN 978-953-307-051-3
Hard cover, 288 pages
Publisher InTech
Published online 01, February, 2010
Published in print edition February, 2010

Human-robot interaction (HRI) is the study of interactions between people (users) and robots. HRI is multidisciplinary with contributions from the fields of human-computer interaction, artificial intelligence, robotics, speech recognition, and social sciences (psychology, cognitive science, anthropology, and human factors). There has been a great deal of work done in the area of human-robot interaction to understand how a human interacts with a computer. However, there has been very little work done in understanding how people interact with robots. For robots becoming our friends, these studies will be required more and more.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Otniel Portillo-Rodriguez, Oscar O. Sandoval-Gonzalez, Carlo Avizzano, Emanuele Ruffaldi and Massimo Bergamasco (2010). Capturing and Training Motor Skills, Human-Robot Interaction, Daisuke Chugo (Ed.), ISBN: 978-953-307-051-3, InTech, Available from: http://www.intechopen.com/books/human-robot-interaction/capturing-and-training-motor-skills

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821